

Least-squares solutions for baselines and matrix factorization

October 14, 2026

1 Setup: training samples and the squared-loss objective

Consider n_u users and n_i items. The training set consists of n observed rating triples

$$\{(u_l, i_l, r_l)\}_{l=1}^n, \quad u_l \in \{1, \dots, n_u\}, \quad i_l \in \{1, \dots, n_i\},$$

meaning that user u_l assigns rating r_l to item i_l . For each user u and each item i , the training histories are the index sets

$$\Omega_u = \{l : u_l = u\}, \quad \Omega_i = \{l : i_l = i\}.$$

We fit a prediction rule $f(u, i)$ by minimizing the *sum of squared errors* (SSE) over the n training samples, possibly together with a regularizer.

Remark 1 (RMSE vs. SSE). The reporting metric is

$$\text{RMSE}(f) = \sqrt{\frac{1}{n} \sum_{l=1}^n (r_l - f(u_l, i_l))^2}.$$

Since $x \mapsto \sqrt{x/n}$ is strictly increasing on $[0, \infty)$ and n is fixed during fitting, minimizing RMSE over any model class is *equivalent* to minimizing the SSE. All fitting objectives below therefore use the plain squared loss; RMSE is only used for reporting on held-out data.

2 Baseline methods: separable least squares

Each baseline below minimizes a squared-loss objective. The organizing device is *separability*: the objective splits into independent one-dimensional problems, each solved by a sample mean.

Theorem 2 (Global Mean RS). *The global mean recommender system solves:*

$$\min_{c \in \mathbb{R}} \sum_{l=1}^n (r_l - c)^2.$$

The optimal solution is

$$\hat{c} = \bar{r} = \frac{1}{n} \sum_{l=1}^n r_l.$$

Then the prediction rule is $\hat{r}_l = \hat{c}$ for all test samples.

Proof. Let $g(c) = \sum_{l=1}^n (r_l - c)^2$, a differentiable function of c . The first-order condition gives

$$g'(c) = -2 \sum_{l=1}^n (r_l - c) = 0 \implies \sum_{l=1}^n r_l = nc \implies c = \bar{r}.$$

Since $g''(c) = 2n > 0$, the objective is strictly convex, so the unique stationary point $c = \bar{r}$ is the unique global minimizer. $\square$

Theorem 3 (User Mean RS). *The user mean recommender system solves:*

$$\min_{a \in \mathbb{R}^{n_u}} \sum_{l=1}^n (r_l - a_{u_l})^2.$$

The optimal solution is, for every user u with $\Omega_u \neq \emptyset$,

$$\hat{a}_u = \bar{r}_u = \frac{1}{|\Omega_u|} \sum_{l \in \Omega_u} r_l.$$

Then the prediction rule is $\hat{r}_l = \hat{a}_{u_l}$ for all test samples; if user u is unseen in training ($\Omega_u = \emptyset$), we fall back to $\hat{a}_u = \hat{c}$.

Proof. Group the training samples by user. The objective *separates* across users:

$$\sum_{l=1}^n (r_l - a_{u_l})^2 = \sum_{u=1}^{n_u} \sum_{l \in \Omega_u} (r_l - a_u)^2,$$

where the inner sum depends only on the single parameter a_u . Hence the joint problem is minimized by minimizing each inner problem independently. For a user with $\Omega_u \neq \emptyset$, the inner problem is exactly the Global Mean RS problem (Theorem 2) over the group $\{r_l\}_{l \in \Omega_u}$, whose unique minimizer is the group mean $\bar{r}_u$. $\square$

Theorem 4 (Item Mean RS). *The item mean recommender system solves:*

$$\min_{b \in \mathbb{R}^{n_i}} \sum_{l=1}^n (r_l - b_{i_l})^2.$$

The optimal solution is, for every item i with $\Omega_i \neq \emptyset$,

$$\hat{b}_i = \bar{r}_i = \frac{1}{|\Omega_i|} \sum_{l \in \Omega_i} r_l.$$

Then the prediction rule is $\hat{r}_l = \hat{b}_{i_l}$ for all test samples; if item i is unseen in training ($\Omega_i = \emptyset$), we fall back to $\hat{b}_i = \hat{c}$.

Proof. Exchange the roles of users and items in the proof of Theorem 3. $\square$

Example 5 (The four-rating table from the slides). Consider the $n = 4$ samples $(0, 0, 5)$, $(0, 1, 1)$, $(1, 0, 1)$, $(1, 1, 5)$. Then $\hat{c} = 3$, $\hat{a}_0 = \hat{a}_1 = 3$, and $\hat{b}_0 = \hat{b}_1 = 3$. Every mean baseline predicts 3 for all four pairs, with training RMSE = 2: additive averages cannot represent that the two users like *opposite* items. This motivates the user–item interaction term introduced in Section 3.

Remark 6 (The three-mean combination). The rule $\hat{r}_l = \frac{1}{3}(\hat{a}_{u_l} + \hat{b}_{i_l} + \hat{c})$ is a *fixed equal-weight combination* of the three fitted estimators above, rather than the minimizer of a joint least-squares problem. Choosing general weights $w_u \hat{a}_{u_l} + w_i \hat{b}_{i_l} + w_0 \hat{c}$ with $w_u + w_i + w_0 = 1$ is therefore a model-selection step (performed on validation data), not a fit on the training set. A genuinely *fitted* additive rule does exist: $f(u, i) = \mu + a_u + b_i$ is the $K = 0$ special case of the matrix-factorization objective in Section 3.

3 Matrix factorization: model and objective

Example 5 showed that no additive average can express a user–item interaction. The (regularized, biased) matrix-factorization model therefore adds an interaction term to the additive baseline:

$$f_{\theta}(u, i) = \mu + a_u + b_i + p_u^{\top} q_i, \quad \theta = (a, b, P, Q),$$

where $\mu = \hat{c}$ is fixed at the global mean (Theorem 2), $a \in \mathbb{R}^{n_u}$, $b \in \mathbb{R}^{n_i}$ are user/item biases, and $P \in \mathbb{R}^{n_u \times K}$, $Q \in \mathbb{R}^{n_i \times K}$ collect the latent factors with rows $p_u^{\top}$ and $q_i^{\top}$. The fitting criterion is

$$\min_{\theta} J(\theta) = \sum_{l=1}^n \left(r_l - \mu - a_{u_l} - b_{i_l} - p_{u_l}^{\top} q_{i_l} \right)^2 + \lambda (\|P\|_F^2 + \|Q\|_F^2). \quad (1)$$

The hyperparameters are the number of factors $K \in \{1, 2, \dots\}$ (chosen before fitting) and the regularization strength $\lambda \geq 0$; biases are deliberately left *unpenalized*.

Remark 7 (The joint problem is nonconvex). The term $p_u^{\top} q_i$ is *bilinear* in the unknowns, so (1) is not jointly convex. Moreover, when $\lambda = 0$ the objective is invariant under the rescaling $(p_u, q_i) \mapsto (\gamma p_u, \gamma^{-1} q_i)$ for any $\gamma \neq 0$, so minimizers are never unique. However, with *one block held fixed*, J is a convex quadratic in the other block. *Alternating least squares* (ALS) exploits exactly this structure: fix one block, optimize the other, and alternate.

Lemma 8 (Block separability). *Fix (a, P) and define the partial residuals $y_l = r_l - \mu - a_{u_l}$. Then*

$$J = \sum_{i=1}^{n_i} \left[\sum_{l \in \Omega_i} (y_l - b_i - p_{u_l}^{\top} q_i)^2 + \lambda \|q_i\|_2^2 \right] + \lambda \|P\|_F^2 = \sum_{i=1}^{n_i} J_i(b_i, q_i) + \text{const},$$

so the item block decomposes into n_i independent subproblems over (b_i, q_i) . Symmetrically, fixing (b, Q) and redefining the partial residuals as $r_l - \mu - b_{i_l}$ decomposes J into n_u independent subproblems over (a_u, p_u) .

Proof. Regroup the training samples by item (resp. by user) and collect the penalty terms that involve the free block. $\square$

4 Each block update is a Ridge regression

Each ALS iteration thus consists of two passes: with (a, P) fixed, solve every item subproblem $J_i(b_i, q_i)$; then with (b, Q) fixed, solve every user subproblem. These subproblems need no new solver—each one is a Ridge regression, and the update can be read off a standard Ridge fit.

Theorem 9 (Item update = Ridge regression). *Fix (μ, a, P) . For item i with $\Omega_i \neq \emptyset$, the subproblem J_i of Lemma 8 is exactly a Ridge regression, with the correspondence*

$$X^{(i)} = [p_{ul}^\top]_{l \in \Omega_i} \in \mathbb{R}^{|\Omega_i| \times K}, \quad y^{(i)} = [r_l - \mu - a_{ul}]_{l \in \Omega_i} \in \mathbb{R}^{|\Omega_i|},$$

$$\text{intercept} = b_i \text{ (unpenalized)}, \quad \text{coefficients} = q_i, \quad \text{penalty strength} = \lambda.$$

That is, the subproblem coincides with

$$\min_{b_i, q_i} \|y^{(i)} - b_i \mathbf{1} - X^{(i)} q_i\|_2^2 + \lambda \|q_i\|_2^2,$$

so the update $(\hat{b}_i, \hat{q}_i)$ is read off the intercept and the coefficient vector of the Ridge fit.

Proof. Ridge regression with an unpenalized intercept minimizes $\|y - b \mathbf{1} - X \beta\|_2^2 + \alpha \|\beta\|_2^2$ over (b, β) . Substituting $X = X^{(i)}$, $y = y^{(i)}$, $\beta = q_i$, $b = b_i$, and $\alpha = \lambda$ reproduces $J_i(b_i, q_i)$ term by term. $\square$

Corollary 10 (User update = Ridge regression). *Fix (μ, b, Q) . For user u with $\Omega_u \neq \emptyset$, the subproblem is exactly a Ridge regression with*

$$X^{(u)} = [q_{il}^\top]_{l \in \Omega_u} \in \mathbb{R}^{|\Omega_u| \times K}, \quad y^{(u)} = [r_l - \mu - b_{il}]_{l \in \Omega_u} \in \mathbb{R}^{|\Omega_u|},$$

intercept a_u (unpenalized), coefficients p_u , and penalty strength λ ; the update $(\hat{a}_u, \hat{p}_u)$ is read off the Ridge fit.

Proof. Swap the roles of users and items in Theorem 9. $\square$

Remark 11 (Why $\lambda > 0$ matters). With $\lambda = 0$ and a short history ($|\Omega_i| \leq K$), the Ridge fit is not unique: the residuals can be made zero by many coefficient vectors. Any $\lambda > 0$ makes the subproblem strictly convex, so the update is well defined and unique no matter how short the history is.

Remark 12 (Empty history $\Rightarrow$ cold-start fallback). If $\Omega_i = \emptyset$, the item subproblem reduces to $\min_q \lambda \|q\|_2^2$, so $\hat{q}_i = 0$ while $\hat{b}_i$ is undetermined; we adopt the convention $\hat{b}_i = 0$. The prediction for a new item is then $\hat{r}_{ui} = \hat{\mu} + \hat{a}_u$. Symmetrically, a new user receives $\hat{\mu} + \hat{b}_i$, and a pair with two new IDs receives $\hat{\mu}$. In short: use whichever blocks are available, and default to $\hat{\mu}$ when none are.

Remark 13 (From baselines to matrix factorization). The mean baselines of Section 2 are nested inside the MF objective. With $K = 0$ the interaction term vanishes and each Ridge update contains only the intercept: Theorem 2 applied to the partial residuals gives, for example,

$$\hat{b}_i = \frac{1}{|\Omega_i|} \sum_{l \in \Omega_i} (r_l - \hat{\mu} - \hat{a}_{u_l}),$$

which, when $a \equiv 0$, is exactly the item-mean baseline recentered at $\hat{\mu}$. The interaction term $p_u^\top q_i$ with $K \geq 1$ is what allows the model to separate pairs that every additive baseline must score identically (Example 5).